心理学报 2021, Vol. 53, No. 9, 1032−1043
Acta Psychologica Sinica

https://doi.org/10.3724/SP.J.1041.2021.01032

基于选项层面的认知诊断非参数方法*

郭 磊[1,2] 周文杰[1]

([1]西南大学心理学部, 重庆 400715) ([2]中国基础教育质量监测协同创新中心西南大学分中心, 重庆 400715)

摘 要 充分挖掘选择题(Multiple-Choice, MC)的诊断信息受到了较多关注, 将干扰项信息考虑在内可以提升诊断精度。为了弥补参数模型基于大样本才能获得可靠估计的不足, 以及适用于班级水平的小样本诊断测验情境, 本研究提出了非参数的多选题诊断方法。模拟和实证研结果表明:(1)当MC测验中题目参数不存在较大差异时, d_{h-MC}法在多数情况下表现优于参数类诊断模型。(2)当MC测验中题目参数存在较大差异时, d_{ph-MC}法的表现最优。(3)实证研究中非参数方法和参数类模型的分类一致性程度较高, d_{ph-MC}距离法估计得到的考生属性总体掌握程度与总分相关最高。最后, 基于MC诊断测验的特点提出了若干研究方向。

关键词 认知诊断评估, 选择题, 干扰项信息, 非参数诊断方法, 汉明距离
分类号 B841

1 引言

心理测验理论在历经长期发展后, 已由标准测验理论过度到新一代测验理论。作为新一代测验理论的认知诊断评估(Cognitive Diagnostic Assessment, CDA), 重在评价学生在知识掌握方面的优劣势, 从而为个性化教学提供依据。因此, 准确估计学生的知识状态(Knowledge State, KS)是个性化教学的重要前提, 若估计不精准, 则补救会有偏差。

目前估计学生KS的方法有两大类, 一类是参数类方法, 一类是非参数方法。前者特点为: 能将被试的潜在能力与题目参数用显式数学表达式, 即认知诊断模型(Cognitive Diagnosis Model, CDMs)呈现, 易于刻画作答时的认知加工过程。但其对总体假设有依赖(如单调性、作答局部独立性、参数不变性等), 且当模型复杂后, 参数估计多采用MCMC算法, 耗时太长, 不适合于小样本的诊断(郭磊 等, 2018; 康春花 等, 2015)。CDM主要包括对数线性模型(LCDM; Henson et al., 2009), 广义DINA模型(GDINA; de la Torre, 2011), 广义诊断模型(GDM; von Davier, 2008), 以及若干简约模型, 如DINA、DINO、A-CDM、RRUM等。后者特点为: 不基于模型, 因此无需对被试总体进行限制, 且估计方便快捷, 尤其适合小样本诊断(Chiu, et al., 2018)。但它无法表征属性间的交互作用, 无法得到模型拟合指标。非参数诊断方法主要有聚类分析法(Chiu, et al., 2009; 康春花 等, 2015; 郭磊 等, 2018)、距离判断法(Chiu & Douglas, 2013; Chiu et al., 2018; 康春花 等, 2019)和机器学习法(李世珍, 2019)等。

不论使用哪类方法, 都需要分析被试在测验上的作答数据才能知晓其KS。当前, 在TIMSS、PISA、NAEP和TOEFL等标准化测验中, 主流题型为选择题(Multiple-Choice, MC), 因为MC题目有如下优势: 不受主观误差影响(Thissen & Wainer, 1993)、能够提高测验信度(Steven, 2004)、易于批阅且计分快速、能够满足内容平衡需求(Osterlind, 1998)等。但目前对MC题目的使用效率较低, 仅对是否选择了正确答案进行评分, 忽略了大量存在于干扰项中的诊断信息(Thissen & Steinberg, 1984; de la Torre, 2009; 李瑜, 2014; 刘拓, 2016)。在数据分

收稿日期: 2020-11-02
* 国家自然科学基金青年项目(31900793); 北京师范大学中国基础教育质量监测协同创新中心重大成果培育性项目(2019-06-023-BZPK01); 中央高校基本科研业务费专项资金(SWU2109222)资助。
通信作者: 郭磊, E-mail: happygl1229@swu.edu.cn

析时未能纳入干扰项信息，不仅是对测验编制的极大浪费，更会降低被试能力的估计精度(Bock, 1972; Thissen, 1976; Levine & Drasgow, 1983; de la Torre, 2009)。为了在分析时纳入干扰项信息，在编制 MC 题目时，不仅需要对正确答案进行编码，还要对干扰项编码。表 1 呈现了一道 4 个选项的分数减法编码例子。该题目考察了 3 个属性：S1 整数借位，S2 分数相减，S3 约分。被试需要掌握 3 个属性才能选择正确答案 D，若只掌握了 S2 属性，会选择选项 A，若只掌握了 S2 和 S3 属性，则会选择选项 B。可以看出，每个干扰项都在起到诊断作用，因此比起二值计分结果，显然会提升诊断精度。

表 1　选项编码的分数减法示例

$2\frac{4}{12}-\frac{7}{12}$		属性		
		S1	S2	S3
A	$2\frac{3}{12}$		√	
B	$2\frac{1}{4}$		√	√
C	$1\frac{9}{12}$	√	√	
D	$1\frac{3}{4}$	√	√	√

注：示例来自 de la Torre (2009, pp.166−167)

为了充分利用 MC 的诊断信息并分析 MC 数据，研究者提出了相应的 MC-CDMs。如, MC-DINA 模型(Multiple-Choice DINA; de la Torre, 2009)，基于选项层面的 SICM 模型(Scaling Individuals and Classifying Misconceptions Model; Bradshaw & Templin, 2014)，多策略的多选题认知诊断模型(李瑜, 2014)，基于多选题选项层面的 GDCM-MC 模型(Generalized Diagnostic Classification Models for Multiple Choice Option-Based Scoring; DiBello, et al., 2015)，以及三个结构化的 MC-DINA 模型(Structured DINA model for multiple-choice items; Ozaki, 2015)。上述模型属于参数类方法，需要在大样本基础上才能获得比较精确的参数估计结果，而且使用到的 MCMC 算法非常耗时。然而，正如 Chiu 等(2018)指出，CDM 更适合用于大规模测验，若将其用于小样本，即在班级水平上监督教学和学习过程中，将得不到准确的参数估计结果。因此，本研究将提出能够分析小样本数据，并且还能充分考虑 MC 题目干扰项信息的非参数诊断方法，旨在最大化 MC 题目的诊断功效，又能适用于小班规模的诊断目标。

2　传统非参分类法简介

如前所述, CDA 中的非参方法主要为聚类分析法、距离判断法和机器学习法。聚类方法的最大不足在于标签识别问题，即无法判断聚类得到的类别的 KS 是哪一种(Chiu et al., 2009; Guo, et al., 2020)。机器学习法的不足在于，该类方法需要提前生成数据对，以生成的数据对来训练神经网络(李世珍, 2019)，数据对的质量很大程度上影响诊断结果，并且该类方法需要消耗大量的计算机算力。为了弥补以上缺点, Chiu 和 Douglas (2013)提出了 3 种基于汉明距离的非参分类(nonparametric classification, NPC)方法，分别是简单汉明距离 d_h，加权汉明距离 d_{wh} 和惩罚汉明距离 d_{ph}。它们分别表述为公式(1)至公式(3)：

$$d_h\left(\boldsymbol{Y}_i,\boldsymbol{\eta}_i\right)=\sum_{j=1}^{J}\left|Y_{ij}-\eta_{ij}\right| \tag{1}$$

其中，$\boldsymbol{Y}_i=\left(\mathrm{Y}_{i1},\mathrm{Y}_{i2},\cdots,\mathrm{Y}_{iJ}\right)$ 为被试 i 在 J 道题目上的实际作答结果，$\boldsymbol{\eta}_i=\left(\eta_{i1},\eta_{i2},\cdots,\eta_{iJ}\right)$ 为被试 i 在 J 道题目上的理想作答结果，有 $\eta_{ij}=\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}$。$K$ 为测验考察的属性个数。α_{ik} 是被试 i 在属性 k 上的掌握情况，掌握时，$\alpha_{ik}=1$，否则 $\alpha_{ik}=0$。q_{jk} 是 Q 矩阵中的元素，表示题目 j 是否考察了属性 k，当考察时，$q_{jk}=1$，否则 $q_{jk}=0$。该方法的判断逻辑是：实际作答与理想作答之间的汉明距离越小，属于该理想作答模式所对应的 KS 的可能性越高，因此可将被试判归到该 KS 中。

$$d_{wh}(\boldsymbol{Y}_i,\boldsymbol{\eta}_i)=\sum_{j=1}^{J}\frac{1}{\overline{p}_j(1-\overline{p}_j)}|Y_{ij}-\eta_{ij}| \tag{2}$$

公式(2)为加权汉明距离，$\overline{p}_j$ 表示题目 j 的实际正确作答比例。若 $\overline{p}_j$ 越大(或越小)，则该题目对答对和答错被试的区分能力就越强，因此其权重就越大。其余符号含义同公式(1)。

$$d_{ph}(\boldsymbol{Y}_i,\boldsymbol{\eta}_i)=\sum_{j=1}^{J}w_g I[Y_{ij}=1]|Y_{ij}-\eta_{ij}|+\sum_{j=1}^{J}w_s I[Y_{ij}=0]|Y_{ij}-\eta_{ij}| \tag{3}$$

公式(3)为惩罚汉明距离，w_g 和 w_s 分别为猜测和失误权重，用以调节不同题型对距离大小的影响(Chiu & Douglas, 2013)，权重的应用场景是当某些题型的猜测行为或失误行为发生的可能性存在较大差异时，则需要对猜测行为或失误行为赋予不同

的权重。如，在开放性题目中，被试几乎不可能在未掌握题目所考察的属性时猜对题目，应给予猜测行为(即对应 $\eta_{ij}=0, Y_{ij}=1$ 时)更多的惩罚。当题目的猜测系数小于失误系数时，有 $w_g > w_s$。当 $w_g = w_s = 1$ 时，惩罚汉明距离等价于简单汉明距离。

随后, Chiu 等(2018)在 NPC 基础上提出了广义分参数分类(general nonparametric classification, GNPC)方法，目的是使非参数认知诊断方法可以更好处理属性之间存在的复杂链接情况。但 GNPC 方法的复杂性更大，且根据研究结果表明，GNPC 和 NPC 的分类准确性在不同实验条件下各有高低，GNPC 未展现出明显优势，因此本研究基于简洁性和实用性考虑，选择在 NPC 基础上开发适用于 MC 题型的非参数方法。

3 MC 测验的非参分类方法

基于传统 NPC 的思想，本研究提出了 3 种可处理干扰项信息的非参数方法。记题目 j 包含 O 个选项，o = 1,2,⋯,O，被试的作答反应向量 $\boldsymbol{Y}_{ij}$ 的结果形式是一个多维向量，Y_{ijo} 是其中的元素，表示被试 i 在题目 j 中选项 o 上的选择情况，若被试选择了选项 o，则 $Y_{ijo}=1$，否则 $Y_{ijo}=0$。MC 测验中被试的理想作答 η_{ijo}^* 可表示为：

$$\eta_{ijo}^* = \eta_{ijo} * \left[1-\prod_{k=1}^{K_j^*}(1-\alpha_{ik})\right] \tag{4}$$

$$\eta_{ijo} = \prod_{k=1}^{K_j^*}[2-2^{(\alpha_{ik}-q_{jok})^2}] \tag{5}$$

K_j^* 表示题目 j 所考察的属性个数，有 $K_j^* \leqslant K$，即将 q 向量中的元素 0 去掉并将元素 1 向前排序，以使这些考察了的属性为前 K_j^* 个属性(de la Torre, 2011)。例如，当 $\boldsymbol{q}_{jo}=(1,0,1)$ 时，$K_j^*=2$，此时，$\boldsymbol{q}_{jo}^*=(q_{jo1},q_{jo3})$，可称为坍缩(collapsed)$q$ 向量，其对应的 KS 即可被称为坍缩 KS。q_{jok} 表示题目 j 在选项 o 上对属性 k 的考察情况，若考察了，$q_{jok}=1$，否则 $q_{jok}=0$。η_{ijo} 用于判断被试 i 在题目 j 上的坍缩 KS 与选项 o 所考察属性之间是否完全一致，$\eta_{ijo}=1$ 表示完全一致，否则 $\eta_{ijo}=0$。η_{ijo}^* 的目的是排除被试随机作答后的理想作答情况，η_{ijo}^* 取值为 1 需要满足两个条件：①$\eta_{ijo}=1$，且有②被试 i 至少掌握题目 j 所考察的一个属性，即有 $1-\prod_{k=1}^{K_j}(1-\alpha_{ik})=1$。由公式(4)可以看出，只有当被试 i 在题目 j 上的坍缩 KS 与选项 o 所考察属性之间完全一致，且使用了真实能力作答时，有 $\eta_{ijo}^*=1$；只要被试 i 在题目 j 上的坍缩 KS 与选项 o 所考察属性之间不完全一致，或者被试 i 随机猜测时，有 $\eta_{ijo}^*=0$。因此，公式(4)可以用来表示被试被选项 o 所“吸引”的程度，即理想作答。

计算实际作答和理想作答之间的汉明距离便可构造分析 MC 题型的非参数诊断方法，分别记作：$d_{h-MC}, d_{wh-MC}, d_{ph-MC}$ 距离。3 种新方法的区别在于，赋予 $\boldsymbol{Y}_i$ 和 $\boldsymbol{\eta}_i^*$ 之间距离的权重方式不同，下面分别介绍 3 种新方法。

3.1 d_{h-MC} 距离法

d_{h-MC} 距离是简单汉明距离 d_h 在 MC 题型上的推广，赋予 Y_{ijo} 与 η_{ijo}^* 中不一致元素相同的权重。该方法中的观测作答 $\boldsymbol{Y}_i$ 和理想作答 $\boldsymbol{\eta}_i^*$ 之间的汉明距离可以定义为：

$$d_{h-MC}(\boldsymbol{Y}_i,\boldsymbol{\eta}_i^*) = \sum_{j=1}^{J}\sum_{o=1}^{O}|Y_{ijo}-\eta_{ijo}^*| \tag{6}$$

这是 3 种新方法中最简洁的一种，对所有 Y_{ijo} 与 η_{ijo}^* 不一致元素的次数进行简单求和，即可得 $\boldsymbol{Y}_i$ 和 $\boldsymbol{\eta}_i^*$ 之间的 d_{h-MC} 距离。

3.2 d_{wh-MC} 距离法

d_{wh-MC} 距离是加权汉明距离 d_{wh} 在 MC 题型上的推广，其表达式为：

$$d_{wh-MC}(\boldsymbol{Y}_i,\boldsymbol{\eta}_i^*) = \sum_{j=1}^{J}\sum_{o=1}^{O}\frac{1}{\bar{p}_{jo}(1-\bar{p}_{jo})}|Y_{ijo}-\eta_{ijo}^*| \tag{7}$$

其中，$\bar{p}_{jo}$ 表示在题目 j 中选择选项 o 的比例，$\frac{1}{\bar{p}_{jo}(1-\bar{p}_{jo})}$ 是实际作答和理想作答在选项水平上的权重。若 $\bar{p}_{jo}$ 越大(或越小)，则该题目对区别选择了选项 o 和未选择选项 o 的被试的能力就越强，因此其权重就越大。特别地，Chiu 和 Douglas (2013)在加权汉明距离中未考虑到在实际测验场景中存在 $\bar{p}_j=0$ 的情况，此时会导致权重的分母为 0。因此，本研究将该情况下的 $\bar{p}_{jo}$ 赋值为较小的常数，如 0.001。

3.3 d_{ph-MC} 距离法

如前文所述，Chiu 和 Douglas (2013)提出的惩罚汉明距离，其本质是对不同猜测和失误行为进行刻画，如，当题目猜测概率很小时(如开放题)，那么发生了猜测行为就应该给予较大的惩罚，即 w_g 应该取较大值。但作者在研究中将惩罚权重 w_g 和 w_s 固定为题目间相同，该设置有个缺陷：当一份测

验中不同题目间质量存在较大差异时，它们的惩罚权重应该不同。因此，将惩罚汉明距离 d_{ph} 推广到 MC 测验中，需要考虑为不同质量的题目设置不同的惩罚权重。d_{ph-MC} 距离可被定义为：

$$d_{ph-MC}(\boldsymbol{Y}_i,\boldsymbol{\eta}_i^*)=\sum_{j=1}^{J}ws_j\left(\sum_{o=1}^{O}I[\gamma_{ij}=1]\,|\,y_{ijo}-\eta_{ijo}^*\,|\right)+\sum_{j=1}^{J}wg_j\left(\sum_{o=1}^{O}I[\gamma_{ij}=0]\,|\,y_{ijo}-\eta_{ijo}^*\,|\right) \tag{8}$$

其中，$\gamma_{ij}=\sum_{o=1}^{O}\eta_{ijo}^*$，用于判断被试 i 是否使用了真实能力作答，仅有被试 i 在题目 j 上的坍缩 KS 完全匹配了题目 j 的某一选项所考察的属性时，γ_{ij} 才等于 1。$I[\cdot]$ 为示性函数，当括号中两者相等，则返回 1，否则返回 0。当 $I[\gamma_{ij}=1]=1$ 的同时发生失误行为，应给予失误权重；而当 $I[\gamma_{ij}=0]=1$ 的同时发生猜测行为时，应给予猜测权重。特别地，当一份测验中题目质量间不存在较大差异时，可固定 $ws_j=wg_j=1$，此时 d_{ph-MC} 距离法等同于 d_{h-MC} 距离法。

4　模拟研究一

4.1　研究目的

本研究拟采用蒙特卡洛模拟方式探讨与诊断模型相比，3 种非参数诊断方法是否能有效提升对被试 KS 的估计精度。

4.2　数据生成方式(真模型)的选择

为了生成作答数据，得使用参数类模型，该做法也是非参诊断研究中常见的做法(Chiu & Douglas, 2013; Chiu et al., 2018)。本研究选取 Ozaki (2015) 所提出的结构化 MC-DINA 模型 1 (MC-S-DINA1, MC1)和结构化 MC-DINA 模型 2 (MC-S-DINA2, MC2)。因为这两个模型比其他 MC-CDMs 更加简洁，根据作者的模拟研究表明，MC1 和 MC2 的诊断精度要优于 MC-DINA 模型(de la Torre, 2009)。未选择 MC3 模型的原因是：它的表现不如 MC2，且比 MC2 更复杂。因此，本研究选择 MC2 作为“最佳情境”的结果参照，MC1 作为“最简洁情境”的结果参照。下面分别对 MC1 和 MC2 模型进行简要介绍。

对于 MC1，被试 i 选择题目 j 的第 o 个选项的概率为：

$$P(Y_{ijo}=1\,|\,\boldsymbol{\alpha}_i)=\gamma_{ij}(1-\delta_j)^{\eta_{ijo}}\left(\frac{\delta_j}{O-1}\right)^{1-\eta_{ijo}}+\frac{(1-\gamma_{ij})}{O} \tag{9}$$

其中，η_{ijo} 的计算方法与公式(5)中的相同，γ_{ij} 的计算方式与公式(8)中的相同。δ_j 是题目的“失误”参数，表示被试没有选择最匹配其 KS 的那个选项的概率，即出现了“失误”。不难看出，MC1 模型中的题目只有一个参数，因此它是最简洁的模型代表。

对于 MC2，被试 i 选择题目 j 的第 o 个选项的概率为：

$$P(Y_{ijo}=1|\,\boldsymbol{\alpha}_i)=\gamma_{ij}(1-\delta_{jo})^{\eta_{ijo}}\left(\frac{\beta_{ij}}{O-1}\right)^{1-\eta_{ijo}}+\frac{(1-\gamma_{ij})}{O} \tag{10}$$

与 MC1 不同之处在于，MC2 的题目参数 δ_{jo} 被定义在选项水平上，因此不同选项具有不同的“失误概率”。$\beta_{ij}=\sum_{o=1}^{O}\delta_{jo}\eta_{ijo}$，表示被试的理想作答取 1 时，而没有选择最匹配选项的概率。当理想作答为 0 时，则公式(10)变为 $\frac{(1-\gamma_{ij})}{O}$，表示完全随机猜测。其余符号含义同公式(10)。

4.3　实验设计

本研究为 4 因素完全交叉设计，4 个自变量分别是样本量($N=30, 50, 100$)、题目长度($J=10, 20, 30$)、题目质量(高质量、低质量)和真模型(MC1, MC2)。在每个实验条件下，分别使用 3 种非参数诊断方法与 2 种诊断模型分析数据，并计算被试的模式/属性判准率。采用与 Ozaki (2015)相同的带有干扰项信息编码的 **Q** 矩阵，共考察 5 个属性，如表 2 所示，表 2 中的数字表示该属性在该题所被考察的次数。以第 23 题 q 向量为例说明，q_{11} = [21000]表明 4 个选项的编码分别为[11000]、[10000]、[00000]和[00000]。当题目长度为 10 时，使用 **Q** 矩阵的后 10 题，当题目长度为 20 时，使用 **Q** 矩阵的后 20 题，题目长度为 30 时，使用整个 **Q** 矩阵。MC1 和 MC2 使用 MCMC 算法进行参数估计，在 R 中实现参数估计，其 MCMC 设置与 Ozaki (2015)相同，且所有参数估计得到的 $\hat{R}$ 值小于 1.1，达到了收敛标准。

被试 KS 真值从多元正态阈值模型(Chiu et al., 2009)中生成，该方法被广泛应用于认知诊断领域中(e.g. Chiu et al., 2009; Chiu & Douglas, 2013; Chiu et al., 2018; Chang et al., 2019)。首先定义一个 K 维向量 $\boldsymbol{\theta}_i=(\theta_{i1},\theta_{i2},\cdots,\theta_{iK})'$ 作为被试 i 在每个属性上的连续能力值，$\boldsymbol{\theta}_i$ 从多元正态分布 $MVN(0,\boldsymbol{\Sigma})$ 中生成，协方差矩阵 $\boldsymbol{\Sigma}$ 的非对角线元素 ρ 可以描述属性间的相关，本研究设置 ρ = 0.5，用以表示中等程度相关(Chiu & Douglas, 2013)，$\boldsymbol{\Sigma}$ 如下所示：

$$\boldsymbol{\Sigma}=\begin{pmatrix}1 & & 0.5\\ & \ddots & \\ 0.5 & & 1\end{pmatrix}$$

表 2　MC 题目中干扰项已编码的 Q 矩阵

属性	题目																													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
A1	1	0	0	0	0	1	0	0	0	0	2	2	1	2	0	0	0	0	0	0	2	1	3	1	2	2	0	0	0	0
A2	0	1	0	0	0	0	1	0	0	0	1	0	0	0	2	1	1	0	0	0	2	2	1	0	0	0	2	2	2	0
A3	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	1	0	1	0	0	2	2	0	2	2	0	2
A4	0	0	0	1	0	0	0	0	1	0	0	0	2	0	0	2	0	2	0	1	0	2	0	2	0	2	2	0	2	2
A5	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	2	0	2	2	0	0	1	0	2	2	0	2	2	2

被试的 KS 真值 $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{iK})$ 可被定义为：

$$\alpha_{ik} = \begin{cases} 1, if\ \theta_{ik} \geqslant \Phi^{-1}\left(\dfrac{k}{K+1}\right); \\ 0, \qquad\qquad otherwise. \end{cases}$$

式中 $\Phi^{-1}(\cdot)$ 是标准正态分布概率密度函数的逆函数，指给定概率值时，所求取的 Z 分数。

当真模型为 MC1 时，高质量题目的参数 δ_j 从 $U(0.1, 0.2)$ 中生成，低质量题目的参数从 $U(0.2, 0.3)$ 中生成。当真模型为 MC2 时，高质量题目的参数 δ_{jo} 从 $U(0.1, 0.2)$ 中生成，低质量题目的参数从 $U(0.2, 0.3)$ 中生成。需要注意的是，在模拟研究一中，由于同一实验条件下的所有题目的题目质量不存在较大差异(在同一分布范围内生成)，故可设定 $d_{ph\text{-}MC}$ 距离法中所有题目惩罚权重均为 1，$d_{ph\text{-}MC}$ 距离法等价于 $d_{h\text{-}MC}$，因此结果仅呈现 $d_{h\text{-}MC}$ 距离法的表现。所有实验循环 100 次以减少随机误差。

4.4　评价指标

使用模式判准率(PCCR)和平均属性判准率(AACCR)评价被试 KS 的估计精度，公式分别如下：

$$AACCR = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} W_{ik}}{N*K} \tag{11}$$

$$PCCR = \frac{\sum_{i=1}^{N}\prod_{k=1}^{k} W_{ik}}{N} \tag{12}$$

其中，当估计得到的 $\hat{\alpha}_{ik}$ 与真值 α_{ik} 相等时，$W_{ik} = 1$，否则 $W_{ik} = 0$。

4.5　结果

表 3 呈现了真模型为 MC1 时，两类诊断方法在

表 3　两类诊断方法的模式判准率和属性判准率(真模型为 MC1)

题目质量	题目数量	样本量	PCCR				AACCR			
			$d_{h\text{-}MC}$	$d_{wh\text{-}MC}$	MC1	MC2	$d_{h\text{-}MC}$	$d_{wh\text{-}MC}$	MC1	MC2
高质量	10	30	**0.784**	0.710	0.763	0.703	**0.918**	0.884	0.906	0.896
		50	**0.783**	0.701	0.749	0.690	**0.916**	0.883	0.900	0.889
		100	**0.789**	0.703	0.757	0.704	**0.922**	0.888	0.902	0.896
	20	30	**0.911**	0.893	0.896	0.888	**0.968**	0.962	0.930	0.928
		50	**0.911**	0.895	0.879	0.863	**0.976**	0.962	0.918	0.970
		100	**0.912**	0.895	0.905	0.896	**0.973**	0.963	0.921	0.968
	30	30	0.957	0.947	**0.979**	0.964	0.987	0.984	**0.992**	0.991
		50	0.951	0.934	**0.973**	0.966	0.986	0.980	**0.992**	0.989
		100	0.954	0.940	**0.976**	0.970	0.986	0.982	**0.993**	0.983
低质量	10	30	**0.575**	0.495	0.498	0.450	**0.843**	0.798	0.814	0.799
		50	**0.588**	0.501	0.505	0.428	**0.843**	0.801	0.820	0.788
		100	**0.590**	0.501	0.518	0.420	**0.849**	0.806	0.828	0.784
	20	30	**0.802**	0.768	0.742	0.655	**0.933**	0.919	0.917	0.888
		50	**0.798**	0.762	0.742	0.651	**0.935**	0.921	0.919	0.889
		100	**0.793**	0.760	0.752	0.671	**0.930**	0.917	0.922	0.892
	30	30	**0.865**	0.849	0.820	0.757	**0.964**	0.959	0.952	0.935
		50	**0.868**	0.845	0.837	0.777	**0.965**	0.957	0.957	0.940
		100	**0.874**	0.853	0.848	0.801	**0.967**	0.959	0.961	0.947

注：粗体表示该条件下的最大值

不同条件下的模式判准率和平均属性判准率。总体上看，在大多数条件下，非参数类方法的 PCCR 和 AACCR 要高于参数类模型。具体来说，第一，题目质量对两类诊断方法均有较大影响，题目质量越高，判准率越高。在高质量题目情况下，d_{h-MC} 在多数情况下表现最优，其 PCCR 的范围为 0.783 至 0.957，AACCR 的范围为 0.918 至 0.987；对于参数类模型而言，MC1 的表现要优于 MC2，其 PCCR 的范围分别为 0.749 至 0.979 以及 0.690 至 0.970 之间，AACCR 的范围分别在 0.900 至 0.993 以及 0.889 至 0.991 之间；d_{wh-MC} 的表现相对较差。在低质量题目的所有情况下，d_{h-MC} 方法的诊断精度能维持在较高水平并且表现最优，其 PCCR 的范围在 0.575 至 0.874 之间，AACCR 的范围在 0.843 至 0.967 之间；d_{wh-MC} 与 MC1 的表现相近，其 PCCR 的范围分别在 0.495 至 0.853 以及 0.498 至 0.848 之间，AACCR 的范围分别在 0.789 至 0.959 以及 0.814 至 0.961 之间，d_{wh-MC} 在 $J = 20$ 和 30 时的 PCCR 高于 MC1，表明在题目质量较低时，若题目数量中等或较多，加权距离法的判准率要优于诊断模型；MC2 在低质量时表现相对较差。该结果表明，题目质量对 MC-CDM 的影响更大，d_{h-MC} 在一定程度上可以缓冲题目质量变低后给诊断精度带来的影响。

第二，题目数量对参数和非参数诊断方法均会带来影响。首先，随着题目数量增多，两类诊断方法的判准率均在提升，但相较而言，MC-CDM 对题目数量的变化更加敏感。例如，在高质量条件下，$J = 10$ 时，d_{h-MC} 表现最优，其 PCCR 在 0.785 左右，AACCR 在 0.919 左右；真模型 MC1 在相同条件下，PCCR 在 0.756 左右，AACCR 在 0.903 左右，存在差距；当题量提升至 20 时，两者仍存在差距，d_{h-MC} 的 PCCR 提升至 0.911 左右，AACCR 提升至 0.967 左右，真模型 MC1 在相同条件下，PCCR 提升至 0.893 左右，AACCR 提升至 0.923 左右；而当题量提升至 30 时，真模型的表现出现反转，d_{h-MC} 的 PCCR 提升至 0.954 左右，AACCR 提升至 0.986 左右，MC1 在相同条件下，PCCR 提升至 0.976 左右，AACCR 提升至 0.992 左右，MC1 表现超过了 d_{h-MC}。而在低质量条件下，无论题目数量如何变化，d_{h-MC} 始终有着最高的判准率，这再次展现了非参数诊断方法处理低题目质量的优势。其次，在题目数量中等及较少条件下，两类诊断方法在判准率上的差距较为明显，但随着题目数量不断增加，差距在不断缩小，特别是在高质量情况下，MC-CDM 的表现会出现反转，这说明题目数量对参数类诊断方法的影响要大于对非参数方法的影响。样本量对两类诊断方法的判准率影响程度较小。

表 4 呈现了真模型为 MC2 时，两类诊断方法在

表 4　两类诊断方法的模式判准率和属性判准率(真模型为 MC2)

题目质量	题目数量	样本量	PCCR				AACCR			
			d_{h-MC}	d_{wh-MC}	MC1	MC2	d_{h-MC}	d_{wh-MC}	MC1	MC2
高质量	10	30	**0.772**	0.700	0.746	0.697	**0.915**	0.884	0.904	0.896
		50	**0.781**	0.700	0.747	0.701	**0.917**	0.880	0.900	0.893
		100	**0.788**	0.705	0.753	0.705	**0.921**	0.889	0.903	0.897
	20	30	**0.907**	0.888	0.887	0.888	0.966	0.961	0.935	**0.967**
		50	**0.909**	0.892	0.884	0.905	0.965	0.959	0.923	**0.972**
		100	0.911	0.896	0.886	**0.916**	0.967	0.961	0.923	**0.971**
	30	30	0.953	0.938	0.960	**0.976**	0.985	0.980	0.991	**0.991**
		50	0.949	0.938	0.966	**0.973**	0.985	0.981	0.989	**0.992**
		100	0.952	0.936	0.972	**0.973**	0.986	0.981	0.987	**0.993**
低质量	10	30	**0.566**	0.501	0.490	0.424	**0.835**	0.798	0.807	0.787
		50	**0.580**	0.493	0.497	0.424	**0.841**	0.797	0.815	0.786
		100	**0.593**	0.501	0.516	0.422	**0.847**	0.803	0.823	0.786
	20	30	**0.787**	0.752	0.723	0.642	**0.931**	0.917	0.915	0.886
		50	**0.793**	0.761	0.744	0.656	**0.930**	0.917	0.917	0.889
		100	**0.792**	0.762	0.754	0.666	**0.931**	0.918	0.921	0.892
	30	30	**0.872**	0.849	0.830	0.759	**0.964**	0.957	0.954	0.935
		50	**0.873**	0.846	0.844	0.777	**0.965**	0.956	0.959	0.940
		100	**0.873**	0.848	0.849	0.797	**0.965**	0.956	0.959	0.945

不同条件下的 PCCR 和 AACCR 结果，其表现与真模型为 MC1 时的结果大体相似。整体上来看，d_{h-MC} 的表现仍是最优。

5 模拟研究二

5.1 研究目的

目前大部分诊断模型研究在探讨题目质量对判准率影响时，均是约束题目参数为固定值，或在同一质量分布内。如 Chiu 等(2018)将题目参数固定为 0.1,0.2 和 0.3，Ma 等(2016)将一份测验中的题目质量约束在 $U(0.05, 0.15)$，或 $U(0.15, 0.25)$ 范围内，这使得一份测验中题目质量过于同质。而在现实情境中，一份测验里的不同题目可能会在质量上存在较大差异。因此，本研究为了贴近现实，拟探讨当一份 MC 测验中不同题目的质量存在较大差异时，非参数诊断方法的表现。

5.2 实验设计

本研究为 3 因素完全交叉设计，3 个自变量分别是样本量($N = 30, 50, 100$)、测验长度($J = 10, 20, 30$)、以及真模型(MC1, MC2)。测验前半部分题目参数 δ_{jo} 从 $U(0, 0.1)$ 中生成，后半部分题目的参数 δ_{jo} 从 $U(0.2, 0.4)$ 中生成。由于前半部分题目的质量较高，当被试在这些题目中发生失误时，将给予更大的惩罚，因此在 d_{ph-MC} 方法中设定前半部分题目的失误权重 ws_j 应更大；而后半部分题目的质量较低，则设定后半部分题目的失误权重 ws_j 更小。由于惩罚权重 ws_j 的设定没有可参考的前人研究作为依据，因此经过多次的预实验探索，本研究最终确定将前半部分题目的失误权重 ws_j 设定为 2，后半部分题目的失误权重 ws_j 设定为 1，猜测权重 wg_j 设定为 1。其余条件与模拟研究一相同。

5.3 实验结果

表 5 呈现了各个诊断方法在不同条件下的模式判准率和平均属性判准率。在所有条件下，d_{ph-MC} 的表现均最优，PCCR 在 0.647 至 0.943 之间，AACCR 在 0.868 至 0.986 之间；d_{h-MC} 与 MC1 的表现次之，当 $J = 10$ 时，d_{h-MC} 表现更好，当题目数量提升至 20 或 30 时，MC1 表现更好，两者的 PCCR 分别在 0.623 至 0.908 以及 0.578 至 0.939 之间，AACCR 分别在 0.820 至 0.970 以及 0.847 至 0.986 之间；d_{wh-MC} 和 MC2 相对表现最差。测验长度大幅提高了两类诊断方法的估计精度，其对参数类模型的提升作用更大，更长的测验长度可以缩小参数类方法与非参数方法的差距。样本量对两类模型的估计精度均有轻微的提高作用。总体而言，当 MC 测验中题目质量存在较大差异时，d_{ph-MC} 对 KS

表 5 题目质量存在较大差异时各方法的模式判准率和属性判准率

真模型	题目数量	样本量	PCCR					AACCR				
			d_{h-MC}	d_{wh-MC}	d_{ph-MC}	MC1	MC2	d_{h-MC}	d_{wh-MC}	d_{ph-MC}	MC1	MC2
MC1	10	30	0.631	0.547	**0.669**	0.596	0.523	0.865	0.820	**0.877**	0.858	0.835
		50	0.644	0.549	**0.675**	0.605	0.518	0.866	0.825	**0.877**	0.856	0.822
		100	0.645	0.543	**0.678**	0.623	0.523	0.869	0.825	**0.880**	0.866	0.826
	20	30	0.839	0.812	**0.888**	0.857	0.796	0.945	0.935	**0.964**	0.958	0.937
		50	0.840	0.817	**0.882**	0.859	0.800	0.948	0.939	**0.964**	0.960	0.938
		100	0.844	0.819	**0.894**	0.877	0.829	0.947	0.937	**0.967**	0.964	0.946
	30	30	0.904	0.878	**0.938**	0.930	0.906	0.975	0.968	**0.986**	0.984	0.978
		50	0.904	0.883	**0.943**	0.933	0.916	0.974	0.968	**0.987**	0.984	0.981
		100	0.908	0.891	**0.942**	0.939	0.925	0.976	0.970	**0.986**	0.986	0.983
MC2	10	30	0.623	0.546	**0.647**	0.578	0.512	0.866	0.820	**0.868**	0.847	0.825
		50	0.638	0.548	**0.672**	0.601	0.521	0.866	0.824	**0.876**	0.858	0.827
		100	0.643	0.548	**0.676**	0.621	0.519	0.870	0.824	**0.879**	0.865	0.825
	20	30	0.834	0.803	**0.886**	0.853	0.801	0.944	0.933	**0.967**	0.957	0.939
		50	0.836	0.808	**0.897**	0.862	0.817	0.942	0.931	**0.969**	0.959	0.944
		100	0.838	0.808	**0.892**	0.868	0.828	0.944	0.932	**0.966**	0.960	0.948
	30	30	0.905	0.879	**0.942**	0.925	0.900	0.973	0.966	**0.986**	0.982	0.976
		50	0.906	0.884	**0.942**	0.928	0.909	0.974	0.968	**0.986**	0.984	0.979
		100	0.905	0.884	**0.937**	0.933	0.924	0.974	0.968	**0.985**	0.984	0.982

表 6　包含干扰项信息的大学英语高级英语阅读测验 Q 矩阵

	题目 1						题目 2						题目 3						题目 4					
A	**1**	**1**	**0**	**1**	**0**	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	**1**	**1**	**0**	**1**	**0**	**1**	0	0	0	0	0	0
D	0	0	0	0	0	0	**1**	**1**	**0**	**1**	**0**	**1**	1	0	0	1	0	0	**1**	**1**	**1**	**0**	**0**	**0**
	题目 5						题目 6						题目 7						题目 8					
A	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
B	**1**	**1**	**1**	**0**	**0**	**0**	**1**	**1**	**0**	**1**	**0**	**0**	1	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	1	0	0	0	0	0	**1**	**1**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**
D	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	题目 9						题目 10						题目 11						题目 12					
A	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	**1**	**1**	**0**	**0**	**0**
D	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**0**	0	0	0	0	0	0
	题目 13						题目 14						题目 15											
A	0	0	0	0	0	0	1	0	0	0	0	0	**1**	**1**	**1**	**0**	**0**	**0**						
B	1	1	0	0	0	0	**1**	**0**	**1**	**0**	**0**	**0**	1	0	1	0	0	0						
C	**1**	**1**	**1**	**0**	**0**	**0**	0	0	0	0	0	0	1	0	0	0	0	0						
D	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0						

注：加粗选项为正确答案选项。

的估计精度最高。

值得注意的是，本研究还发现：即使真模型为 MC2，使用 MC1 估计得到的判准率也要高于 MC2 估计的结果，与之前研究呈现出不一样的结果，这说明当测验中的题目质量异质(即差异较大)时，MC1 的表现更加稳定。

6　实证研究

6.1　数据说明

本研究实证数据来自某高校大学英语高级英语阅读考试中的 15 道选择题，被试量为 607 人。考虑到需要将干扰项信息包含在 **Q** 矩阵中，因此，我们邀请了 4 名外国语学院教授(其中 2 位参与了编题工作)对 15 道题目的所有选项进行了编码。具体而言，首先，采用文献分析法，根据《语言文字规范(GF 0018-2018)》中对中国英语学习者英语阅读能力的界定，结合本测验考查内容，共析出了 6 个属性，分别是：A1：提取细节；A2：理解句子间关系；A3：推测隐含意义；A4：概括主旨要义；A5：推断作者情感态度；A6：理解修辞手法。4 名教授分别独立标定，之后计算他们在所有选项上标定的一致性，即肯德尔 W 系数，得到 $W = 0.938$, $p < 0.001$，表明 **Q** 矩阵标定的一致性较高。**Q** 矩阵如表 6 所示。

分别使用两类诊断方法对实证数据进行分析。为了评估非参数诊断方法和 MC-CDM 的表现，参考 Chiu 等(2018)的做法：①计算非参数方法与 MC-CDM 的分类一致性，包括平均属性分类一致性指标($AAR = \dfrac{\sum_{i=1}^{N}\sum_{k=1}^{K} I[\hat{\alpha}_{ik}^{1} = \hat{\alpha}_{ik}^{2}]}{N*K}$ ，其中上角标 1 表示由第 1 种方法得到的估计值，上角标 2 表示由第 2 种方法得到的估计值)，模式分类一致性指标 1 ($PAR(K=6) = \dfrac{\sum_{i=1}^{n} I[\hat{\alpha}_{i}^{1} = \hat{\alpha}_{i}^{2}]}{N*K}$)，以及模式分类一致性指标 2($PAR(K \geqslant 5) = \dfrac{\sum_{i=1}^{N} I\left[\sum_{k=1}^{K} I[\hat{\alpha}_{ik}^{1} = \hat{\alpha}_{ik}^{2}] \geqslant 5\right]}{N*K}$)。②计算两类诊断方法估计得到的属性总体掌握程度[1]与总分间的相关。由于 $d_{\text{wh-MC}}$ 在两个模拟研究中均表现不佳，所以在实证研究中剔除了 $d_{\text{wh-MC}}$ 的结果。

[1]属性总体掌握程度是将被试的属性掌握模式(即 KS)进行求和，如[11100]的被试的掌握程度记为 3。

6.2 研究结果

表 7 展示了各模型间的三种分类一致性程度。首先，同类型的诊断方法之间的分类一致性程度更高，非参数方法与 MC-CDM 之间的分类一致性程度相对较低但仍处于较高水平。具体而言，MC1 与 MC2 之间的 AAR, PAR (K = 6)以及 PAR (K ≥ 5)分别为 0.92, 0.71 和 0.94, $d_{\text{h-MC}}$ 与 $d_{\text{ph-MC}}$ 之间的 AAR, PAR (K = 6)以及 PAR (K ≥ 5)分别为 0.88, 0.61 和 0.92, 非参数方法与 MC-CDM 的 AAR 在 0.84~0.86 的范围内, PAR (K = 6)在 0.51~0.59 的范围内, PAR (K ≥ 5)在 0.87~0.89 的范围内。其次，相比于 $d_{\text{h-MC}}$, $d_{\text{ph-MC}}$ 与 MC-CDM 的分类一致性程度更高；相比于 MC2, MC1 与非参数方法的分类一致性程度更高，但差异较小。总的来说，各诊断方法之间的分类一致性程度较好，而同类型的诊断方法间的一致性程度更高。

另外，还计算了两类诊断方法估计得到的属性总体掌握程度与总分间的相关。其中，$d_{\text{ph-MC}}$ 为 0.779, MC1 为 0.745, $d_{\text{h-MC}}$ 为 0.743, MC2 为 0.740。可以看出，$d_{\text{ph-MC}}$ 表现最好，而 MC1、$d_{\text{h-MC}}$ 和 MC2 的表现稍差，且三者之间的差异很小。综合而言，考虑到该测验属性个数较多(K=6)而测验长度较短(J=15)，各诊断方法之间的分类一致性程度在可接受范围内。

7 讨论与研究结论

7.1 讨论与展望

在模拟研究一中，当题目数量为 30 题且题目质量高时，真模型的效果好于非参方法，而在 20 题和 10 题时，真模型效果均差于非参方法。如表 3 所示，MC1 为真模型，样本量为 50，题目质量为低时，测验长度从 30 题增加到 20 题时，$d_{\text{h-MC}}$ 方法的 PCCR 下降了 0.070，而 MC1 下降了 0.095，测验长度从 20 题下降至 10 题，$d_{\text{h-MC}}$ 方法的 PCCR 下降了 0.210，而 MC1 下降了 0.237；固定测验长度为 20 题时，题目质量从高下降至低时，$d_{\text{h-MC}}$ 方法的 PCCR 下降了 0.113，而 MC1 下降了 0.137。在题目质量较差或测验长度较短时，非参数方法的效果优于真模型，其可能的原因是测验长度和题目质量对诊断模型的影响更大，而对非参数方法的影响更小。例如，在 Chiu 等(2018)中，以 DINA 模型为真模型时，在样本量为 50，属性数为 5，题目质量中等的条件下，当测验长度从 50 题下降至 30 题时，NPC 方法的 PCCR 下降了 0.150, DINA 模型下降了 0.170；在固定 30 题时，题目质量从高下降至中时，NPC 方法的 PCCR 下降了 0.230, DINA 模型下降了 0.290。我们推测，在 MC 测验中，当题目质量较差或测验长度较短时，干扰项对非参数方法的效果提升高于对诊断方法的效果，从而使得非参数方法能在题目质量较差或测验长度较短时表现更佳。综上所述，$d_{\text{h-MC}}$ 方法在题目质量较差或测验长度较短时具有较强的稳健性。

此外，在研究一中发现，当 MC2 为真模型，在高题目质量中的 20 题和 30 题条件下，相比于 MC1, MC2 的表现更好；而在高质量题目中的 10 题条件下，MC1 的表现更好；低质量题目的所有条件下，MC1 的表现也要更好。其原因可能在于, MC2 对题目质量和测验长度的敏感性更高，只有当题目质量较高，测验长度较长时，MC2 才能有较好的表现。因此，在模拟研究二中，由于一半题目的质量较差，导致即使以 MC2 为真模型，其表现也差于 MC1，这说明当测验中的题目质量异质(即差异较大)时，MC1 的表现更加稳定。

Chiu 和 Douglas (2013)提出的加权汉明距离方法相较于简单汉明距离方法表现稍好，而本研究基于加权汉明距离逻辑提出的 $d_{\text{wh-MC}}$ 在 MC 测验中的表现不及简单汉明距离 $d_{\text{h-MC}}$，这表明可能不能直接将传统的加权汉明距离推广到 MC 测验的干扰项层面，需要结合 MC 题目的特点，提出更合理的权重计算公式，以提高加权汉明距离在 MC 测验中

表 7 各模型间的分类一致性程度

指标	平均属性分类一致性指标 (AAR)				模式分类一致性指标 1 (PAR(K = 6))				模式分类一致性指标 2 (PAR(K ≥ 5))			
	$d_{\text{h-MC}}$	$d_{\text{ph-MC}}$	MC1	MC2	$d_{\text{h-MC}}$	$d_{\text{ph-MC}}$	MC1	MC2	$d_{\text{h-MC}}$	$d_{\text{ph-MC}}$	MC1	MC2
$d_{\text{h-MC}}$	1				1				1			
$d_{\text{ph-MC}}$	0.88	1			0.61	1			0.92	1		
MC1	0.85	0.86	1		0.55	0.59	1		0.88	0.89	1	
MC2	0.84	0.85	0.92	1	0.51	0.57	0.71	1	0.87	0.88	0.94	1

的适用性。另外, Chiu 和 Douglas (2013)提出的惩罚权重汉明距离方法, 以及本文针对 MC 测验提出的 $d_{\text{ph-MC}}$ 在没有题目先验信息的情况下, 惩罚权重的取值需要通过预实验来确定, 目前尚无更好的方法来确定惩罚权重。例如, 在本文的实证研究中, 首先根据 MC1 估计得到 15 道题目的失误参数将 15 道题划分为高、中、低三个质量区间。随后将低质量题目的失误权重 w_s 设置为基准值 1, 中等质量题目设置失误权重 w_s 为 $1+X$ (X 为正), 高质量题目设置失误权重 w_s 为 $1+X+Y$ (Y 为正)。由于题目的猜测概率都为 $\frac{1}{O}$, 所以可以设置所有题目的猜测系数 w_g 为常数。通过调整 X 和 Y 的值, 然后带入 $d_{\text{ph-MC}}$ 的公式进行估计, 再计算被试的估计属性掌握程度与总分的相关, 当相关达到最高时, 此时的失误权重 w_s 则为最优值。需要指出的是, X 和 Y 的取值在不同的测验间需要根据上述步骤进行调整, 以适配相应的测验情景。此外, 未来可以进一步探讨更合理、更简便、更一般化的惩罚权重设置方法而无需通过预实验方法来确定惩罚权重。

当前 MC 测验中, 对 Q 矩阵界定的限制过强。首先, 在 de la Torre (2009)提出的 MC-DINA 框架下, 需要约束干扰项的 q 向量编码是正确答案 q 向量的子集, 但在实际编制测验时, 会限制干扰项的编码空间。如, 当题目包含多个答题策略时(不同策略的 q 向量可能不一样), 那么干扰项的 q 向量就可以采用另一个策略中的属性组合方式进行编码; 或是当诊断测验考察了迷失概念(misconceptions)时, 干扰项的 q 向量可以设计成考察这些迷失概念而非认知属性(Bradshaw & Templin, 2014)。其次, Ozaki (2015)提出的 MC-S-DINA 限制同一题目中干扰项的 q 向量不能重复使用, 而在实际编制测验时, 时常出现此情况, 不方便选项编码。未来可以在考虑放松 MC 测验 Q 矩阵限制的前提下, 提出更一般化的 MC 诊断模型及非参数的诊断方法。

最后, 已有研究者(Yigit et al., 2019)开发出了基于 MC 测验的认知诊断计算机自适应测试(MC-CD-CAT), 考虑干扰项信息的 CD-CAT 可以实现仅用很短的测验就能显著提高考生分类的精度。CD-CAT 需要基于大样本进行参数校准后才能得到较为精确的题目参数, 从而保证被试能力估计的精度。若基于小样本校准, 得到的题目参数质量较差, 此时基于模型的能力估计精度就无法保证。而非参数 CD-CAT 可以有效提高在小样本情景下的能力估计精度(Chang et al., 2019), 更适合在班级水平使用, 所以结合干扰项信息的非参数 CD-CAT 值得研究, 实现同时兼顾小样本规模, 短测验长度, 高判准精度的目的。

7.2 研究结论

本研究提出了 3 种非参数的 MC 诊断方法, 基于模拟和实证研究结果, 得出如下结论:

(1)相比于 MC-CDM, 非参数 MC 诊断方法在大多数实验条件下表现更优秀, 判准率更高, 尤其在题目质量较差, 测验长度较短时效果更好。此时, 推荐使用简单汉明 d_{h-MC}。

(2)当整个测验中不同题目质量存在较大差异时, 惩罚权重汉明 d_{ph-MC} 的表现最好, 考虑优先使用。

(3)与 Chiu 和 Douglas (2013)的结果不同, 加权汉明 d_{wh-MC} 的结果在三种非参数 MC 诊断方法表现最差, 加权汉明距离不适用于直接推广到 MC 测验中。

(4)在实证数据分析中, 非参数类诊断方法与 MC-CDM 估计得到的被试属性掌握情况的一致性程度较高。并且, 由估计的被试属性总体掌握程度与其总分的相关结果表明, 带惩罚系数的汉明距离得到的相关最高, 因此可知 d_{ph-MC} 表现最好。d_{h-MC} 与两种 MC-CDM 的表现相当。

参　考　文　献

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: a psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*(3), 403–425.

Chang, Y.-P., Chiu, C.-Y., & Tsai, R.-C. (2019). Nonparametric CAT for CD in educational settings with small samples. *Applied Psychological Measurement, 43*(7), 543–561.

Chiu, C.-Y., & Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*(2), 225– 250.

Chiu, C.-Y., Douglas, J. A., & Li, X. D. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*(4), 633–665.

Chiu, C.-Y., Sun, Y., & Bian, Y. H. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika, 83*(2), 355–375.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*(3), 163–183.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A

family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement, 39*(1), 62–79.

Guo, L., Yang, J., & Song, N. Q. (2018). Application of spectral clustering algorithm under various attribute hierarchical structures for cognitive diagnostic assessment. *Journal of Psychological Science, 41*(3), 735–742.

[郭磊，杨静，宋乃庆. (2018). 谱聚类算法在不同属性层级结构诊断评估中的应用. *心理科学, 41*(3), 735–742.]

Guo, L., Yang, J., & Song, N. Q. (2020). Spectral clustering algorithm for cognitive diagnostic assessment. *Fronties in Psychology*, *11,* 944. doi: 10.3389/fpsyg.2020.00944

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210.

Kang, C. H., Ren, P., & Zeng, P. F. (2015). Nonparametric cognitive diagnosis: A cluster diagnostic method based on grade response items. *Acta Psychologica Sinica, 47*(8), 1077–1088.

[康春花，任平，曾平飞. (2015). 非参数认知诊断方法：多级评分的聚类分析. *心理学报, 47*(8), 1077–1088.]

Kang, C. H., Yang, Y. K., & Zeng, P. F. (2019). Approach to cognitive diagnosis: The manhattan distance discriminating method. *Journal of Psychological Science, 42*(2), 455–462.

[康春花，杨亚坤，曾平飞. (2019). 一种混合计分的非参数认知诊断方法：曼哈顿距离判别法. *心理科学, 42*(2), 455–462.]

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43*(3), 675–685.

Li, S. Z. (2019). *Application of back propagation neural network based teaching cognitive diagnosis* (Unpublished master's thesis). Henan Normal University, China.

[李世珍. (2019). *基于 BP 神经网络的教学认知诊断方法及应用* (硕士学位论文). 河南师范大学.]

Li, Y. (2014). *The construction for cognitive diagnosis tests of multiple-choice items and the development of multiple-choice cognitive diagnosis model for multiple strategies*. (Unpublished doctoral dissertation). Jiangxi Normal University, China.

[李瑜. (2014). *多选题认知诊断测验编制及多策略的多选题认知诊断模型的开发* (博士学位论文). 江西师范大学.]

Liu, T. (2016). *Using distractor information in computerized adaptive testing* (Unpublished doctoral dissertation). Beijing Normal University.

[刘拓. (2016). *干扰项信息在计算机化自适应测验中的利用* (博士学位论文). 北京师范大学.]

Ma, W. C., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, Model selection, and attribute classification. *Applied Psychological Measurement, 40*(3), 200–271.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance and other formats* (2nd ed.). Boston: Kluwer Academic.

Ozaki, K. (2015). DINA Models for multiple-choice items with few parameters: Considering incorrect answers. *Applied Psychological Measurement, 39*(6), 431–447.

Steven, M. D. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*(9), 1006–1012.

Thissen, D. M. (1976). Information in wrong responses to the raven progressive matrices. *Journal of Educational Measurement, 13*(3), 201–214.

Thissen, D. M., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49*(4), 501–519.

Thissen, D. M., & Wainer, H. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103–118.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307.

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. *Applied Psychological Measurement*, *43*(5), 388–401.

Nonparametric methods for cognitive diagnosis to multiple-choice test items

GUO Lei[1,2], ZHOU Wenjie[1]

([1] *Faculty of Psychology, Southwest University, Chongqing* 400715, *China*)

([2] *Southwest University Branch, Collaborative Innovation Center of Assessment toward Basic Education Quality, Chongqing* 400715, *China*)

Abstract

Cognitive diagnostic assessment (CDA) focuses on evaluating students' advantages and disadvantages in knowledge mastering, providing an opportunity for individualized teaching. Therefore, CDA has attracted attention of many scholars, teachers, and students at domestic and overseas. In CDA and a large number of standardized tests, multiple-choice (MC) are typical item types, which have the advantages of not being affected by subjective errors, improving test reliability, being easy to review, scoring quickly, and meeting the needs of content balance. To fulfil the potential of MC items for CDA, researchers proposed the MC-cognitive diagnosis models (MC-CDMs). However, these MC-CDMs pertain to parameter methods, which need a large sample size to obtain accurate parameter estimation. They are not suitable for small samples at class level, and the MCMC algorithm is very time-consuming. In this study, three nonparametric MC cognitive diagnosis methods based on hamming-distance are proposed, aiming at maximizing the diagnostic efficacy of MC items and being suitable

for the diagnosis target of a small sample.

Simulation study 1 considered four factors: sample size (30, 50, 100), test length (10, 20, 30), item quality (high and low), and the true model (MC-S-DINA1, MC-S-DINA2). Three nonparametric MC methods and two parametric models were compared. The results showed that in most conditions, the pattern accuracy rates and average attribute accuracy rates of the nonparametric MC method($d_{\mathrm{h-MC}}$) were higher than those of parametric models, especially when the test length was short or item quality was low.

In a real test situation, the quality of different items in a test may vary greatly. Based on this, simulation study 2 set the first half of the items at high quality and the remaining items at low quality. The results showed that the pattern accuracy rates and average attribute accuracy rates of the nonparametric MC method ($d_{\mathrm{ph-MC}}$) were higher than those of the parametric models in all conditions.

In an empirical study, the nonparametric MC methods and the parametric models were used to analyze a set of real data simultaneously. The results showed that nonparametric MC methods and parametric models presented high classification consistency rates. Furthermore, the $d_{\mathrm{ph-MC}}$ method had satisfactory estimations.

In sum, $d_{\mathrm{h-MC}}$ was suitable in most conditions, especially when the test length was short or the item quality was low When the quality of different items was quite diverse, $d_{\mathrm{ph-MC}}$ was a better choice compared with parameteric approaches.

Key words cognitive diagnostic assessment, multiple-choice item, distractor information, nonparametric diagnostic method, hamming distance